Research on Fuzzy Clustering Algorithms for Large Dimensional Data Sets Under Cloud Computing

Shuang-cheng Jia(✉) and Feng-ping Yang

Alibaba Network Technology Co., Ltd., Beijing 100102, China
xindine30@163.com

Abstract. With the rapid development of modern computer technology, the internet technology based on computer technology has also been remarkably developed, which has made great progress in modern information technology. The management and effective application of these large-scale data has become the main trend of the development of modern society. In order to improve the ability of fast processing and recognition of large data, data clustering analysis is needed. Aiming at the problems existing in traditional fuzzy clustering algorithm, this paper proposes a design of fuzzy clustering algorithm for high-dimensional and large data sets under cloud computing. Through data classification and data processing classification, the design of fuzzy clustering algorithm for high-dimensional large data sets is realized, and compared with the traditional algorithm through experiments. The simulation results show that the method is easy to calculate and fast in the fuzzy clustering of high-dimensional and large data sets. The clustering effect is good, and the clustering algorithm of high-dimensional and large data sets is well realized.

Keywords: Computer technology · Cluster analysis · Fuzzy clustering algorithm · High-dimensional large data sets

1 Introduction

With the rapid development of Internet technology, data storage and data compression technology, the emergence of interactive applications such as micro-blog, micro-messaging, social networks, the rise of cloud applications, and the wide use of various forms of digital devices, the data is explosive growth. All sectors of society, such as academia, enterprises and government departments, have paid close attention to the issue of big data. Data has become a new asset, which can bring endless social and economic benefits [1]. Big data is gradually becoming a powerful tool for people to understand and transform the world, making it easier for people to grasp the laws of things and accurately predict the future.

The resources in cloud computing are usually virtualized resources provided through the Internet [2]. There are many high-dimensional data sets in cloud computing. At present, there are still many doubts and disputes about the basic concept and key technologies of high-dimensional large data. Traditional relational database technology is not competent for the processing of these data, because the starting point of

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2021
Published by Springer Nature Switzerland AG 2021. All Rights Reserved
S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 347, pp. 295–305, 2021.
https://doi.org/10.1007/978-3-030-67871-5_27

relational database system is to pursue data consistency and fault tolerance. Diversity is an important feature of large data, which means the universality of data sources and the complexity of data types. In this complex data environment, processing large data is a great challenge. To deal with large data, it is necessary to integrate data from all data sources, extract relationships and entities from them, and store these data in a unified structure after association and integration. Therefore, the pretreatment of large data is one of the key factors to determine the value of large data. In order to reflect the value of large data, the pretreatment of large data needs further processing [3]. Fuzzy clustering is one of the key technologies of large data analysis and an important method to reflect the value of large data [4].

At present, relevant personnel have proposed a large number of data clustering methods, among which, reference [5] proposes an algorithm to increase the confidence threshold. The algorithm repeats the constant confidence threshold algorithm several times while increasing the confidence threshold. The algorithm can almost certainly converge to a certain equilibrium point. A numerical example is given to illustrate the applicability of this method to clustering problems, but the calculation time of the method is slow. Reference [6] proposed a new error aware data clustering (EDC) technology in cluster head (CHS) application. Histogram based data clustering (HDC) module groups time-related data into clusters and eliminates relevant data from each cluster. Recursive outlier detection and smoothing (rods) of HDC module provides error aware data clustering, which uses temporal correlation of data to detect random anomalies. The experimental results show that the algorithm has a small amount of calculation and can reduce a large number of redundant data with the minimum error, but the data coverage is not high.

To solve the above problems, this paper proposes a fuzzy clustering algorithm for high-dimensional large data sets in cloud computing.

2 Design of Fuzzy Clustering Algorithms for High Dimensional and Large Data Sets

2.1 Data Object Classification

Before realizing the fuzzy clustering algorithm, the data obtained should be classified, the original data set should be divided into several data fragments, and these data fragments should be copied to the node of the task, and the data at the node should be processed separately. The main task of this process is to calculate the degree of membership of data objects to clustering centers in nodes. In the process of data classification, the data objects and initial clustering centers (or the clustering centers updated in the last iteration process) at this node are input, the index numbers of the clustering centers are output, and the membership degree of all data objects and data objects in the node are also given. Because the data generated in the classification process are stored on local disks, and these data can be merged locally, in order to reduce the communication overhead between the nodes and the calculation of the process. Map process is used in classification, which is carried out at the node, and can play the role of merging data at this node. The main task of map process is to calculate

the sum of the product of all the corresponding subordinate degrees and the corresponding data objects, and to represent the sum of all subordinate degrees (fuzzy factors) for each cluster center. The map process outputs in the form of seu, which represents the index of the clustering center. The output is the data set composed of the index of the data center, the data object and the membership degree. The data set cooperates as the input of the map process. The input form is to compose a data set with the same values in the data set generated in the stage, so the essence of the process is to correspond to each value. A clustering center is indexed, so a clustering result can be obtained every time a process is executed.

After the above work is completed, the data are classified. In the correlation analysis of data classification, the corresponding data classification methods are different when the types of attribute values of data objects are different. The values of attributes are divided into continuous variables and classification variables. Interval variables or ratio variables are classified by the values of continuous variables. If the data object has the largest degree of membership to a clustering center, then the data object belongs to the corresponding class of the clustering center and classifies the data object according to the data parameters obtained above. The structure of the data object is shown in Fig. 1.

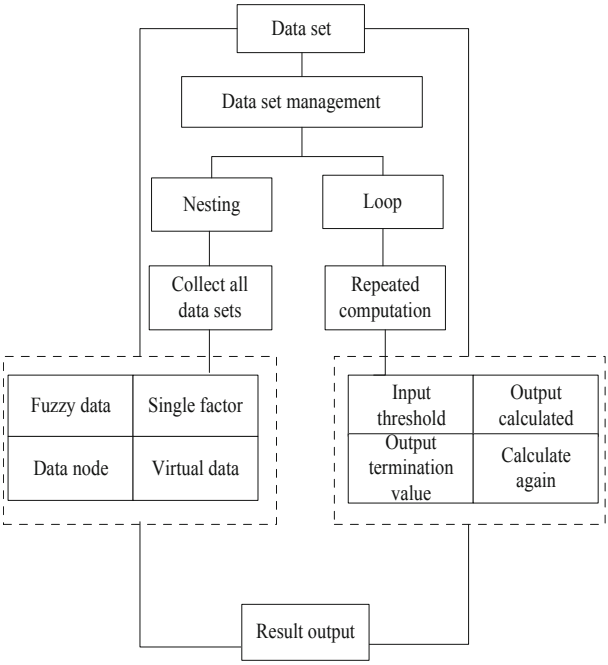


Fig. 1. Data object classification structural diagram.

The structure chart of data object classification shown in Fig. 1, when classifying data, classifies the data according to the process shown in the figure to ensure the accuracy of classification. The classification formula is as follows:

$$f_i = \sum_{d}^{2} vb \frac{x_i - 1}{\max(x_i) - \min(x_m)} \tag{1}$$

In the formula, f_i represents the marked data; $x_i - 1$ represents the standard value of data evaluation; max represents x_i types of data; min represents m types of data, this calculation does not do directional analysis.

$$j = \frac{F_0 \cdot \omega(f_i)}{F \cdot v(j_0 + 1)} \tag{2}$$

In the formula, j represents the classified data after discretization; f represents the standard value of classification; F_0 represents the detected data; j_0 represents the quality level of the data being tested, this calculation does not do directional analysis.

After the above formulas, data object classification is realized. On the basis of object data classification, data processing is carried out on large and high-dimensional data sets. Through this step, the calculation steps of fuzzy clustering can be simplified and the accuracy of fuzzy clustering algorithm can be strengthened.

2.2 Data Processing

Data processing includes two aspects: data extraction and data loading. The main task of data extraction is to obtain data structures and data from the database, and then convert them into corresponding files. Data extraction includes data acquisition and data processing [7]. Firstly, through the analysis of the user's needs to obtain data, select the data table and corresponding data fields, then use database operation technology to connect the corresponding database system, and finally use appropriate statements to load the structure information and content of the data table into the data processing program. After the data is selected, the data is processed, and the data loaded by the data acquisition step is pretreated. Then these data information and the structure of database tables are transformed into documents (or strings) and the interface for users to obtain information is provided [8], which realizes cross-language and cross-platform data acquisition and provides seamless connection for data source acquisition in data loading stage (Fig. 2).

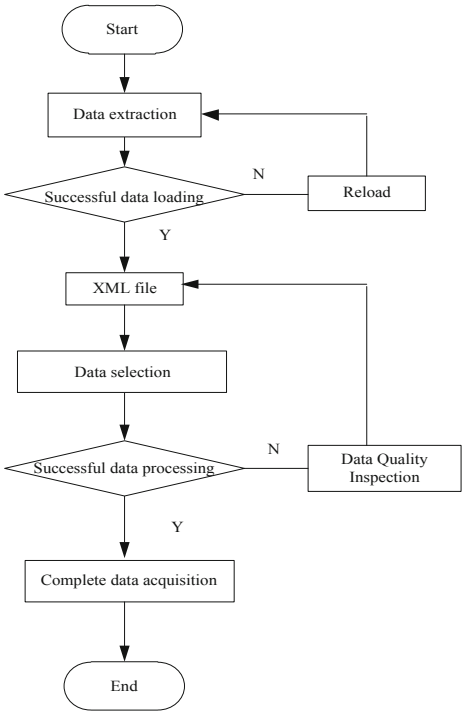


Fig. 2. Data processing diagram.

Data extraction is the basis of realizing the fuzzy clustering algorithm. Through data extraction, we can prepare for the next data loading. After obtaining the file, we need to analyze the XML file, and then create the appropriate parsing method according to the description of the XML and the table structure we need to store. According to the research of document parsing based on technology used by nodes in the platform, data can be processed in parallel. The acquired files are very small and can be read from the table as the input path of the whole job or read in parallel as the process input at the stage of setting up the job.

After data extraction, data is loaded. In this stage, the main task is to access different Web Service interfaces, parse the acquired XML data and load it to the target data source. In this paper, Base is used as the target data source [9]. The address and service method of each data source are loaded into it. By distributing these information to different nodes, each node accesses different data sources according to the information received by each node, and then stores these documents into the file system. At the same time, the information of these files (including storage address and file) is stored. Size, etc.) is stored in the database for subsequent management of these files. When data acquisition is completed, these documents need to be parsed, so we need to look at the preamble of these documents to understand the storage mode of each document and the format of the original data, and then design the corresponding conversion rules and store them. The paths and corresponding transformation rules are

obtained from the database system and processed by adding them to different nodes. The acquired data are finally integrated and stored in the database to provide the basis for the completion of the fuzzy clustering algorithm [10].

3 Implementation of Fuzzy Clustering Algorithms for High Dimensional and Large Data Sets

Fuzzy clustering algorithm is used to describe the uncertain relationship among patterns. It has the characteristics of self-learning, self-organization and self-adaptation. This algorithm has strong robustness and is suitable for parallel processing. Through data preparation and data object analysis, the specific algorithm for designing the fuzzy clustering algorithm is as follows:

Detection data and data classification set are initialized. Set $X = \left(x_{1,}x_{2,}x_3 \ldots x_m\right)$ is assumed to be all the sample points to be clustered, and the number of clustering results of all sample points is c. Sample points are divided into c categories. The data set calculates the membership matrix U from the sample points to the clustering centers iteratively, and then updates the membership matrix U to calculate the new clustering center V, so that the clustering objective function F (U, V) value reaches the minimum, and the clustering results obtained at this time have the highest accuracy [11–13]. The objective function of the fuzzy algorithm can be expressed as:

$$F(U,V) = \min C/\lambda_n f \in X \tag{3}$$

In the formula, λ_n represents the sample size; C represents the fuzzy index; $\in$ represents the clustering effect; X represents the number of iterations of clustering data, this calculation does not do directional analysis.

According to the membership matrix U, which can be obtained from the above formula, the new clustering center is calculated according to the membership matrix U:

$$F = f_i + \sum_{i=1}^{k} A + M \cdot K \tag{4}$$

The detailed steps of the fuzzy clustering algorithm are as follows:

Initialization algorithm stop threshold, the number of iterations is $i = 1$, the maximum number of iterations is K;

Randomly select k sample points as the initial clustering center, calculate the membership degree of each sample point to each initial clustering center, and obtain a membership matrix F;

The clustering centers are recalculated according to the membership degree F, and all kinds of clustering centers are updated [14]. The fuzzy clustering algorithm terminates and outputs the clustering results.

From the implementation steps of the fuzzy algorithm, it can be seen that calculating the membership degree from a sample point to each clustering center is the most important work, and also the most time-consuming step. Since the membership degree

of each sample point is independent of each other, the parallel execution of this process can be considered [15, 16]. It is possible to compute the membership degree of cluster centers of sample points in parallel. By calculating the membership degree of each sample point in parallel on distributed cloud platform with high-dimensional large database, the time required to compute the membership matrix U can be greatly shortened, thus greatly improving the operation efficiency of the fuzzy algorithm.

4 Experimental Demonstration

In order to verify the validity and feasibility of the proposed fuzzy clustering algorithm, a simulation platform is built and experimental analysis is carried out. At the same time, the algorithm is debugged in the context of cloud computing to maximize the accuracy of the experimental results. In order to ensure the rigor of the experiment, the traditional fuzzy clustering algorithm is compared with the fuzzy clustering algorithm designed in this paper. Because there are many kinds of high-dimensional and large data, the elements to be detected are complex, and the data obtained are generated in a short period and high frequency, the data stability is not high. In order to evaluate the quality of clustering results, it is necessary to define the evaluation criteria of clustering and the objective function or method of clustering by checking the clustering results of the algorithm to determine whether the algorithm needs further iteration.

4.1 Experimental Environment and Data Setting

Taking My-sea database as the data source, 10 data sets are selected from the database for experiment, and data analysis is conducted through online data analysis software MOA (an experimental tool for massive online analysis). Table 1 shows the data set used in the experiment.

Table 1. Parameters of experimental data set.

Dataset number	Data contained	Data dimension	Subseries
1	512	4	15
2	1024	12	10
3	261	8	12
4	258	6	16
5	637	10	10
6	298	8	8
7	186	6	12
8	254	12	14
9	301	10	15
10	209	4	5

Table 2 shows the parameters of computer virtual simulation platform.

Table 2. Parameters of computer virtual simulation platform.

Name	Parameter
CPU	Intel i7-8700, 3.75 GHz
Hard disk	512G Solid state disk
Graphics card	NVIDIA
Memory	16G
Operating system	Windows 8.1

According to the above experimental conditions, the data clustering results of different methods are compared, and the experimental conclusions are obtained.

4.2 Clustering Effect Comparison

In the implementation of clustering algorithm, the sum of squares of errors is used as the condition of clustering convergence, and the precision and recall rate are used to evaluate the quality of clustering results. The precision ratio refers to the ratio of the number of data objects correctly clustered into the current cluster and all objects clustered into the current cluster after clustering. Recall ratio refers to the ratio of the number of objects accurately clustered into the current cluster to the number of actual objects in the cluster after clustering.

1000 feature vectors are extracted from high-dimensional large database. In order to reflect the effectiveness of the proposed fuzzy data algorithm and traditional algorithm in computing, the same cluster and the same data set are used for clustering. The results are shown in the Table 3.

Table 3. Traditional algorithm result table.

Data set	Average precision	Average recall rate	Iteration times	Execution time	pH
64	0.39	4.5	12	90	8.1
83	0.42	3.6	10	89	8.3
52	0.52	2.1	11	74	8.2
178	0.69	5.7	12	98	8.2
124	0.78	6.1	14	100	8.1
281	0.96	2.5	11	114	8.1

Setting the same data set, the average precision rate is set to 0.5, the average recall rate is 5.5, the normal number of iterations is 5, the standard value of execution time is set to 60, and the pH is set to 7.

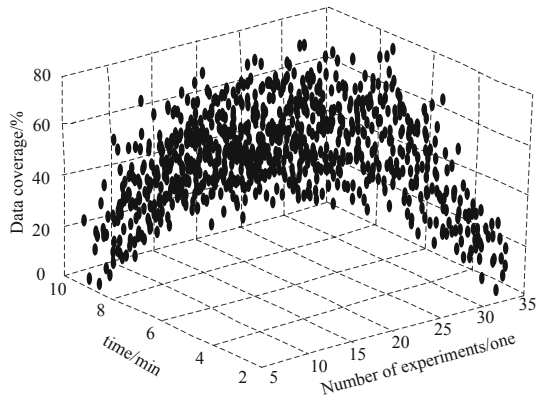
Cloud computing is used to record the results of the fuzzy clustering algorithm, and the average precision, recall, iteration times, execution time and PH of the two algorithms are compared according to the experimental data. The analysis of Table 3 and Table 4 shows that the average precision of the fuzzy clustering algorithm designed in this paper is similar to the standard values, and the average recall rate is higher than that of the traditional algorithm, which has obvious advantages. In terms of execution time, it costs less time and saves a lot of manpower and financial resources. Therefore, we can confirm the validity of the fuzzy clustering algorithm designed in this paper. When clustering high-dimensional large data sets, the number of detection iterations is low, and the operation time is short. It can accurately and quickly cluster high-dimensional large data sets. To a certain extent, it improves the clustering effect of large data, and in a sense, it has the promotion effect. This is because the fuzzy clustering algorithm is used to describe the uncertain relationship between patterns. The algorithm has the characteristics of self-learning, self-organization and self-adaptive. It has strong robustness and is suitable for parallel processing, thus improving the clustering effect of high-dimensional large data sets.

Table 4. The result table of the algorithm in this paper.

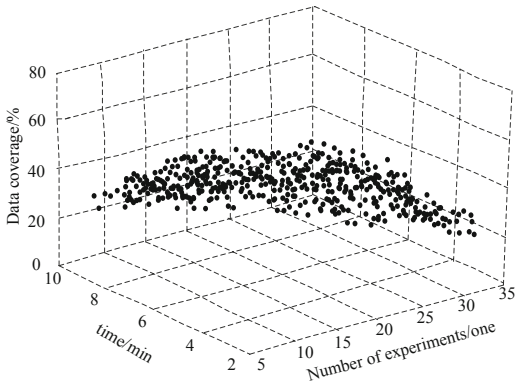
Data set	Average precision	Average recall rate	Iteration times	Execution time	pH
64	0.42	4.6	1	52	8.1
83	0.5	5.1	2	45	8.3
52	0.51	5.2	2	56	8.2
178	0.58	5.6	3	60	8.2
124	0.48	5.9	4	61	8.1
281	0.52	6.0	1	63	8.1

In order to compare the data clustering results of different methods more intuitively, the clustering results of different methods are presented in the form of image comparison. The results are shown in Fig. 3.

Analysis of Fig. 3 shows that the data coverage of the algorithm proposed in this paper is significantly higher than that of the traditional algorithm, the maximum data coverage of traditional algorithm is less than 40%, while the maximum data coverage of this algorithm is more than 70%, indicating that the clustering effect of the algorithm in this paper is better. This is because the algorithm realizes the design of fuzzy clustering algorithm for high-dimensional large data sets through data classification and data processing classification, and effectively improves the data coverage.



(a)The result table of the algorithm in this paper



(b)Traditional algorithm result

Fig. 3. Data clustering results of different methods.

5 Concluding Remarks

With the rapid development of information science and technology and mobile Internet technology, people's life is full of various information data, which is undoubtedly exponential explosive growth. The era of big data has come quietly, and it promotes the development of all walks of life. In order to fully explore the development potential of various industries behind information and data, the high-dimensional and large data fuzzy clustering algorithm under cloud computing has become an indispensable tool. Many existing clustering methods have some problems. Through the analysis of the problems, the design of the fuzzy clustering algorithm is proposed. The experiment proves the effectiveness of the proposed algorithm. The algorithm can cluster high-dimensional large data sets accurately, reduce the residual rate of the fuzzy algorithm, and has very high effectiveness. But there is still room for further study.

References

1. Madadi, S., Mohammadi-Ivatloo, B., Tohidi, S.: A data clustering based probabilistic power flow method for AC/VSC-MTDC. IEEE Syst. J. **13**(4), 4324–4334 (2019)
2. Pengfei, L., Chunyu, L.N.: Real-time flow controllable clustering algorithm for large correlated data in cloud computing environment. Sci. Technol. Eng. **18**(7), 185–190 (2018)
3. Yecheng, G.: Prediction strength optimization of large data clustering algorithms based on distributed computing. J. Changchun Normal Univ. **37**(6), 71–73 (2018)
4. Kuwil, F.H., Shaar, F., Topcu, A.E., et al.: A new data clustering algorithm based on critical distance methodology. Expert Syst. Appl. **129**(9), 296–310 (2019)
5. Nguyen, L.T.H., Wada, T., Masubuchi, I., et al.: Bounded confidence gossip algorithms for opinion formation and data clustering. IEEE Trans. Autom. Control **64**(3), 1150–1155 (2019)
6. Alam, M.K., Aziz, A.A., Latif, S.A., et al.: Error-aware data clustering for in-network data reduction in wireless sensor networks. Sensors **20**(4), 1011 (2020)
7. Tadie, A.T., Guo, Z.: Optimal planning of grid scale PHES through characteristics-based large scale data clustering and emission constrained optimization. Energies **12**(11), 2137 (2019)
8. Bacciu, D., Castellana, D.: Bayesian mixtures of hidden tree markov models for structured data clustering. Neurocomputing **342**(21), 49–59 (2019)
9. Xiaodong, Z.: Research on parallel scheduling algorithms for large data of ships based on cloud computing. Ship Science and Technology **23**(4), 31–33 (2018)
10. Zheng, Q., Diao, X., Cao, J., et al.: From whole to part: reference-based representation for clustering categorical data. IEEE Trans. Neural Netw. Learn. Syst. **31**(3), 927–937 (2020)
11. Kaimin, L., Zili, Z.: Research on the connotation and solutions of big data mining based on cloud computing. Comput. Fans **91**(04), 39 (2018)
12. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)
13. Liu, S., Glowatz, M., Zappatore, M., Gao, H., Jia, B., Bucciero, A. (eds.): eLEOT 2018. LNICST, vol. 243, pp. 1–374. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93719-9
14. Ben HajKacem, M.A., N'cir, C.-E.B., Essoussi, N.: One-pass MapReduce-based clustering method for mixed large scale data. J. Intell. Inf. Syst. **52**(3), 619–636 (2017). https://doi.org/10.1007/s10844-017-0472-5
15. Liu, S., Lu, M., Li, H., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genet. **10**, 120 (2019)
16. Wenying, Z., Junwei, Z., Binning, L., et al.: Research on higher vocational education resource sharing based on cloud computing and big data. Comput. Educ. **22**(04), 12–56 (2018)